UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

May 22, 2012

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-12

MEMORANDUM FOR David C. Whitford
Chief, Decennial Statistical Studies Division

From: Patrick J. Cantwell *(Signed)*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Douglas Olson
Robert Sands
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Report: Net Coverage Comparison with Post-stratification

This report is one of twelve documents providing estimation results from the 2010 Census Coverage Measurement (CCM) program. This report compares estimates of person coverage from the 2010 CCM to those that would have been produced under a post-stratification methodology.

For more information contact Doug Olson on (301) 763-9290.

Attachments

cc:
DSSD CCM Contacts List

Census Coverage Measurement Estimation Report

Net Coverage Comparison with Post-stratification

Prepared by
Douglas Olson
Robert Sands

Decennial Statistical Studies Division

Table of Contents

Executive Summary

This document summarizes the net coverage measurement of persons by the 2010 Census Coverage Measurement program, as contrasted with estimates calculated using post-stratification methods like those employed in the 2000 Accuracy and Coverage Evaluation. The Census Coverage Measurement produced, as in the past, net coverage estimates showing undercount or overcount results. This document does *not* make comparisons to 2000 results, but compares differences due to the use of different *methods* in those two programs.

The 2010 Census Coverage Measurement estimated a net census overcount of 36 thousand persons, or 0.01% (0.14% standard error) in 2010. Use of an analogous post-stratification would have estimated an overcount of 410 thousand persons, or 0.14% (0.13% standard error). Although this difference is not statistically significant, it demonstrates what is probably improvement in methodology through the use of modeling, because increases in estimated undercount (with only a small increase in standard error) usually imply that the capturing of differences in estimated rates was improved.

The Census Coverage Measurement methodology using logistic regression demonstrated significant improvements in the estimation of certain hard-to-count populations whose undercounts have been modeled inadequately using traditional post-stratification, such as persons in Update/Enumerate areas and household residents who are not part of the householder’s nuclear family. The improvements of coverage measurement through modeling should help to indicate populations and operations deserving of efforts for improvement in future censuses.

1. Introduction

The purpose of the 2010 Census Coverage Measurement (CCM) program is to evaluate coverage of the 2010 Census and to improve future censuses. The CCM is designed to measure the census coverage of housing units and persons, excluding group quarters and persons residing in group quarters. The CCM uses a probability sample of 170,000 housing units in the United States. Remote areas of Alaska are out of scope for the CCM. The CCM program provides estimates of net coverage and components of census coverage by using a post-enumeration survey.

This report compares the net coverage estimates for persons in housing units using modeling methods that were employed for the first time in the 2010 CCM, to the results that would have been obtained using post-stratification. The 2000 Accuracy and Coverage Evaluation (A.C.E.) used post-stratification to create its estimates, as had prior coverage measurement evaluations. This report does *not* compare the 2010 CCM to the 2000 A.C.E. results, but compares results using different *methods* on the same underlying data from the 2010 CCM. It also does not compare estimation of housing units because of large differences between the characteristics used for the 2000 post-stratification and those of the 2010 CCM models.

2. Methods

In this section, we discuss briefly the estimation method used in generating the net coverage for persons. For more details on the CCM estimation methodology, see Mule (2008).

2.1 Dual System Estimation

Since the 1950 census, the Census Bureau has been conducting post-enumeration evaluations to estimate the size of error in census counts for areas and demographic groups and to use the information to improve census processes. The post-enumeration survey for 2010, called the 2010 CCM survey, relied on dual system estimation (DSE) that requires two independent systems of measurement. The Population Sample, P sample, and the Enumeration Sample, E sample, have traditionally defined the samples for dual system estimation. The P sample and the E sample measure the same housing unit and household population. However, the P-sample operations are conducted independent of the census. The E sample consists of census housing units and person enumerations in housing units in the same sample areas as the P sample. After matching with the census lists and reconciliation, the P sample provides information about the population missed in the census, whereas the E sample provides information about erroneous census inclusions. This information is used in different ways to estimate the net coverage and the components of census coverage.

For 2010, instead of the post-stratification previously used for coverage estimates, we employ logistic regression modeling to estimate the parameters in the DSE formula for correct enumeration and match probabilities. The model uses all the characteristics from the 2000 A.C.E., as well as some new ones described in Section 2.5, but the interactions between characteristics in the model are in many cases different from the ones used in the post-stratification. We then estimate net error by comparing the estimate of the true population (from the DSE) to the census count, resulting in either a net undercount or a net overcount. The DSE can be expressed as:

$$DSE = \sum_{j \in C} \pi_{dd(j)} \times \frac{\pi_{ce(j)}}{\pi_{m(j)}} \times CB_j$$

With respect to the given estimation domain C, the modeled correct enumeration, match, and data defined (DD) probabilities for census case j ($\pi_{ce(j)}$, $\pi_{m(j),}$ $\pi_{dd(j)}$) are obtained for 2010 through logistic regression modeling. In 2000, they were obtained from post-stratification, in which each census person j, as well as each E sample and P sample person, were assigned to exactly one post-stratum within which the rates for CE, Match, and DD were tabulated among the members. See Olson (2012) for more details on the logistic regression models used to compute the correct enumeration, match, and data-defined probabilities in the above DSE formula.

2.2 Net Coverage Estimates

The comparisons across methods in this report emphasize percent net coverage estimates. The percent net error is the net error estimate divided by the DSE expressed as a percentage.

$$Percent\ \ Net\ \ Undercount = \left(\frac{DSE - Census}{DSE} \right) \times 100$$

2.3 Statistical Testing

Statements of comparison between CCM estimates in this report are statistically significant at the 90% confidence level ($\alpha = 0.10$) using a two-sided test. “Statistically significant” means that the difference is not likely due to random variance from sampling alone, but ignores biases that might arise from non-random causes like synthetic bias or model selection error.

2.4 Post-stratification

The post-stratification applied in the 2000 A.C.E. used six characteristics although not all of the populations were partitioned using all six (U.S. Census Bureau 2004). All of those characteristics are used in the 2010 CCM logistic regression modeling, although some have been defined in a new way. Post-stratification is a special case of logistic regression modeling, in which every observation can be assigned exactly one indicator covariate. The post-stratification used for comparison in this research generally defines characteristics consistently with their 2010 definitions, although it forms groups similarly to the 2000 A.C.E. Here is a summary of the A.C.E. post-stratification variables, with a brief overview of their use in CCM. If the 2010 CCM had used post-stratification instead of modeling, the post-strata would likely have been defined similarly to this way. (Race/Hispanic Origin Domain is described in Mulligan (2012). “TEA/MSA size” refers to the Metropolitan Statistical Area (MSA) size crossed d with Type of Enumeration Area (TEA).)

Table 1: Census 2000 A.C.E. Post-stratification Variables

Characteristic	A.C.E.	CCM	This Research
Race/Hispanic Origin Domain	7 groups	Same	Same
Tenure: Owner or Renter	2 groups	Same	Same
Sex and Age	8 groups	9 groups	9 groups
Tract Return Rate	2 groups	Continuous*	Same as A.C.E.*
MSA size/TEA	4 groups	7 groups	Same as A.C.E.
Region	4 groups	Same	Same

* CCM uses Tract Participation Rate instead of Return Rate

In Table 1, the first three characteristics were the primary stratifiers applied to all persons, with the additional characteristics used if the post-stratum sample size was large enough to support it. The ability to bifurcate the post-stratum was primarily a function of the size of the Race/Origin Domain. Additionally, any post-stratum that included fewer than 100 P-sample members was collapsed across adult ages within sex (collapsing of children was never necessary in 2000.)

Table 2: Summary of the A.C.E. Post-stratification Groups and its Application to this Research

Race/ Origin Domain	Tenure	Return/ Partcptn Rate	MSA/ TEA	Region	Age/ Sex	A.C.E. Post-strata w/ Collapsing		This Research w/ Collapsing	
						Pre	After	Pre	After
White	Owners	2	4	4	8/9*	256	240	288	213
	Renters	2	4		8/9*	64	64	72	71
Black	2	2	2		8/9*	64	60	72	72
Hispanic	2	2	2		8/9*	64	60	72	72
Asian	2	2 (2010 Only)			8/9*	16	16	36	36
NHPI[1]	2				8/9*	16	8	18	18
AIAN[2] off Res	2				8/9*	16	16	18	18
AIAN[2] on Res	2				8/9*	16	16	18	18

*Eight Age/Sex categories were used in 2000 before collapsing, nine in 2010

[1] NHPI = Native Hawaiian and Pacific Islander

[2] AIAN = American Indian and Alaska Native

In Table 2, the first four characteristic columns define the groups that were then sub-divided by age and sex. There were eight Age/Sex categories used in 2000 before collapsing, and nine in 2010. Children have been split into three groups 0-4, 5-9, 10-17, instead of the two used in 2000, which combined the younger groups. Other than that, the only group re-definition is that Asians have adequate sample size to partition by Participation Rate.

In 2000, each post-stratum was permitted a minimum size of 100 P-sample persons. If a post-stratum did not contain that many members all adult age groups of the same sex were collapsed together. For the 2010 comparison, a minimum size of 75 was required due to the smaller sample. Where that size was not met, Region was the first order of collapsing. This resulted in the loss of 75 post-strata of White Owners. No other collapsing was necessary, except in one group of White Renters we had to collapse Children 0-4 and 5-9 together.

2.5 Clustering: An alternative Post-stratification

One of the advantages logistic regression modeling affords over the use of post-stratification is the ability to incorporate additional covariates into the model that the latter could not support. In the 2010 CCM, the characteristics were the following:

- Relationship to Householder: Nuclear member, adult child, other household member.
- Presence of Spouse: Whether the household contains a member with relationship spouse.
- Replacement Mailing or Bilingual Block: Blocks in which enhanced census operations were conducted (these are described in Appendix 2: Census Operational Areas).

Attempting to include these characteristics within the A.C.E. post-stratification would have been infeasible because some post-strata would contain too few sample cases. Since part of the advantage modeling has over post-stratification is the ability to incorporate additional characteristics to reflect heterogeneity, it would be desirable to design a post-stratification scheme that was able to incorporate those differences. Hence, an alternative method has been designed, solely for purposes of researching the effect of the additional variables.

The experimental post-stratification is based on "clustering," which is the partitioning of a population into groups based on the similarity of their characteristics. To do this, the SAS program FastClus was used to construct a partitioning of persons based on the similarity of their modeling characteristics. Two children living in the same housing unit who are members of the same Race/Origin Domain and differ only in that one is eight years old and the other is 12 years old, differ from each other only to the extent their ages motivate different estimates of their modeled CE, Match, and DD rates. Two people of the same age, sex, and Race/Origin Domain in the same household, one the householder and the other a roommate, will differ in the measured distinction between their householder relationships.

To implement clustering, a main-effects logistic regression model estimated the CE and Match rates, using main effects only. The difference between two people is the gross difference between their parameters in the two models, measured using Euclidean distance. See Appendix 1 for technical details about the operation of FastClus. A summary of the kind of partitions it created is in Section 2.7 below.

Because of the environment in which CCM data are studied, some additional controls were necessary to force the algorithmic application of FastClus to fit with the other procedures of the CCM estimation system:

1. Because of the application of Correlation Bias Adjustment factors, the sexes had to be kept separate. Children, Males 18+, and Females 18+ are never in the same partition.
2. Because of the emphasis on studying Race/Origin Domain and child age groups, their parameters were raised to greatly reduce, but not entirely eliminate, the possibility of being joined into a single group together.
3. The American Indians and Alaska Natives (AIAN) on Reservation group was partitioned separately from others.

2.6 Correlation Bias Adjustment

In addition to the data-defined, correct enumeration, and match rates, population estimates reflect a correlation bias adjustment that is applied to adult males only. It is estimated from sex ratios derived from demographic analysis (DA), as compared to the ratio estimated by CCM:

$$c_k = \frac{r_{DA,k}}{r_{PREDSE,k}}$$

where $r_{DA,k}$ and $r_{PREDSE,k}$ are the adjusted DA and Preliminary (i.e., from CE, Match, and DD only) sex ratios, respectively, for age-race group k. As a result of this adjustment, the final DSE sex ratio equals the adapted DA sex ratio within each adult age-race group. There is no correlation bias adjustment for children ages 0 to 17 and adult females. Therefore, the correlation bias adjustment factor for these groups is one. The groups k within which the calculation is performed are the adult age groups 18-29, 30-49 and 50+, divided into Black alone-or-in-combination and Non-Black. These groups do not change based on the post-stratification scheme employed. No ratio could be less than one, which defaulted the 18-29 Non-Black rate for CCM to that value. Table 3 lists the ratios applied in this research.

Table 3: Correlation Bias Adjustment Factors

Race and Age		CCM	Post-stratification	Clustering
Black	18-29	1.0512	1.0566	1.0498
	30-49	1.1111	1.1149	1.0944
	50+	1.0544	1.0552	1.0593
Non-Black	18-29	1.0000	1.0000	1.0008
	30-49	1.0277	1.0287	1.0269
	50+	1.0163	1.0165	1.0163

2.7 Sample Partitions formed by Post-stratification

The tables in the two attachments present the results of the two partitioning methods, post-stratification and clustering, on the creation of estimation groups.

Table A-1 in Attachment A illustrates the way partitions were formed under the two grouping schemes. To read the table, the first line of data shows that AIAN on Reservations represent 0.19% of the weighted E sample; under the A.C.E. post-stratification, the average E-sample person with that characteristic was put into a post-stratum consisting 100% of people with the same characteristic; the next column shows that under Clustering, AIAN on Reservations were also put into groups consisting 100% of themselves. (It was designed to treat AIAN on Reservations separately from other populations.) Looking at the primary A.C.E. post-stratification characteristics, one sees in the two columns describing the A.C.E. post-stratifications that all the primary stratifiers put people into groups consisting 100% of people with those characteristics, with the minor exception that the two youngest ages of children had to be collapsed together in one case, resulting in a group very slightly below 100%. The clustering method enhanced the importance of the Race/Origin Domains. So it generally clustered persons in the same Domain together, except that the Domains for Hispanics and Blacks were intermingled, reflecting that the clustering methodology did not see their parameters as being very different.

Looking at the secondary A.C.E. post-stratification characteristics (the ones between the two blank lines in the center of the table), there are measurements of grouping that in some cases are in the high 90s such as Mailout/Mailback, and others that while not as high are still grouped together much more so than they would be randomly. The Northeast for example contributed 18% of the weighted sample, but were partitioned together over 64% of the time, largely because Region was a primary stratifier among white owners. The characteristics that were not used in the A.C.E. post-stratification (those at the bottom of the table) still show some tendency to group together, because these characteristics are sometimes correlated with others. For example, adult children (looking at the household relationship group at the bottom of the table) are only about 8% of the E sample, but were partitioned into groups consisting of over 36% of that characteristic, because this population group tends to be concentrated among 18 to 29-year-olds. The Clustering on the other hand, considered adult children to be a very distinct group and clustered them into partitions consisting 97% of themselves.

The cost of clustering some of these distinct characteristics together is that other characteristics that had been strongly grouped under A.C.E. post-stratification are less so under clustering. The adult ages, those 18 and over, had been 100% partitioned under A.C.E. post-stratification, but are only in the 90's for 18 to 29-year-olds, and 80's for the older groups.

2.8 Sample Rates Estimated by Different Methods

Tables B-1 and 2 in Attachment B show the effect that the different modeling schemes have on the rates applied to sample persons with each characteristic. It is a mathematical property of logistic regression that the modeled values of a categorical characteristic will equal the average observed rate for that characteristic in the sample. These tables show that the observed and logistic modeling average CE and Match rates for all the categorical characteristics are identical. The Participation Rate was modeled continuously, so does not identically equal its observed rate. The primary A.C.E. stratifiers also identically equal their observed rates, as they must under a post-stratification that never combines them into groups with other observations. The secondary A.C.E. post-stratification characteristics (Region of the country, Mailout/Mailback and the Participation Rate groups) show good correspondence between their post-stratified and observed rates, with the differences generally moving in the direction of the national average rates of 91.80% correct enumerations and 91.07% matches.

The characteristics that were not used in the A.C.E. post-stratification, the Bilingual, Blanketed and Targeted mailing areas, and the householder relationships, were applied rates by post-stratification that in some cases reflected their observed values poorly. The observed match rate in the Update/Enumerate areas for example, was 84.02% and estimated by post-stratification at 87.54%, because under the post-stratification scheme it was combined with Update/Leave as a single characteristic. The match rate of Other Household Members was estimated at 88.93% under post-stratification, against an observed rate of 83.07%. The correct enumeration rate for adult children was observed at 87.77% and estimated at 89.63%.

The clustering partition was constructed to capture these distinctions and generally did so well. The aforementioned CE rate for Adult Children was applied at 87.74% under clustering, very similar to its observed rate. The match rate for Update/Enumerate areas was also nearly identical to its observed rate at 84.03%. These are characteristics that had been clustered strongly (as shown in Attachment B), that had not been partitioned together well under the post-stratification. The purpose of the clustering method was to group these characteristics together to apply their distinctive rates, which was not possible under the A.C.E. post-stratification due to small sample size.

The cost of grouping these characteristics together under clustering is that some of the other groups now have applied rates that are different from those applied under post-stratification and logistic modeling. Since the adult age groups had been primary post-stratification variables, but were clustered with other ages, their applied rates under clustering have been compromised in the direction of the national averages. For example, the low match rates for 18 to 29 Males and Females are now both being applied somewhat higher than their observed rates, while Males and Females over 50 are being applied at lower rates.

This shows the inevitable choices, priorities, and compromises that are necessary in designing a post-stratification. Some characteristics have to be treated as more important than others, and the latter will have their rates smoothed out in their application. The logistic regression modeling however captures all of these differentiations, applying to every modeled characteristic its appropriate rate.

Table 4: Variation among Modeled Values of Sample Cases

Method	Mean	Std Dev	Minimum	Maximum
Correct Enumeration				
Observed	0.9181	0.2628	0.0000	1.0000
Logistic Regression	0.9181	0.0436	0.6171	0.9887
ACE Post-stratified	0.9181	0.0369	0.7856	0.9830
Clustering	0.9181	0.0420	0.7429	0.9686
Match				
Observed	0.9107	0.2797	0.0000	1.0000
Logistic Regression	0.9107	0.0583	0.3792	0.9976
ACE Post-stratified	0.9107	0.0500	0.6748	0.9929
Clustering	0.9107	0.0560	0.6471	0.9714

Logistic regression modeled the E sample as having a standard error of 4.36% among the estimated (modeled) CE rates of the sample observations. A.C.E. post-stratification captured about 85% of that variation at 3.69%, and Clustering 96% at 4.20%. Logistic regression modeled the P sample as having a standard error of 5.83% among estimated match rates; A.C.E. post-stratification captured 86% of that variation, and Clustering 96%. Assuming a model has not been over-fitted for variation that is purely random, an increased internal variance among modeled values should reflect that correlation bias has been reduced through better capture of differences among rates, assuming the model has been correctly specified.

3. Limitations

Certain limitations to the data need to be noted when reading this document.

3.1 Sampling Error

Since the CCM estimates are based on a sample survey, they are subject to sampling error. As a result, the sample estimates will differ from what would have been obtained if all housing unit persons had been included in the survey. The standard errors provided with the data reflect mainly variations due to sampling. They do not in general account for nonsampling errors which can be the principal source of error for very small geographic areas. Thus, the standard errors provide an indication of the minimum amount of error present in the estimates. See the forthcoming methodology report for more details on variance estimation.

3.2 Nonsampling Error

Nonsampling error is a catch-all term for errors that are not a function of selecting a sample. They include errors that may occur during data collection and processing survey data. For example, while an interview is in progress, the respondent might make an error answering a question, or the interviewer might make an error asking a question or recording the answer. Sometimes interviews fail to take place or households provide incomplete data. Other examples of nonsampling error for the 2010 CCM include modeling error, synthetic error, and classification error. Unlike sampling error, nonsampling error is difficult to quantify.

4. Results

Section 4.1 compares national totals using the methodologies described in Section 2 to those obtained from the CCM logistic regression. Sections 4.2 and 4.3 compare net coverage estimates for important demographic groups and persons showing other characteristics.

4.1 Brief Overview Comparison of Partitioning Schemes

To illustrate the effect that the use of increasingly complete models has on DSE’s, coverage estimates using two highly simplified post-stratification schemes from the CCM sample were calculated, before the application of correlation bias adjustment. These are not intended or presented as serious efforts at constructing alternative coverage measurements.

A one-cell DSE treats the entire population as a single post-stratum. The 18-cell calculation uses only the nine age/sex groups, partitioned into Black and Non-Black. As each partitioning includes more characteristics than the previous one, the DSE increases as (in principle) greater amounts of heterogeneity of capture probabilities by the two systems are reflected in the models.

Table 5: Some Coverage Estimates Without and With Correlation Bias Adjustment (thousands)

Partitioning Method	Without Correlation Bias Adjustment	After Correlation Bias Adjustment
Census Count*	300,703	
One Cell	297,105	
18 Cells	297,372	
Post-stratification	297,661	300,293
Clustering	297,930	300,412
CCM	298,111	300,739

The standard error of all estimates is about 410-417 thousand.
*Census count excludes persons in group quarters and Remote Alaska

4.2 Net Coverage for Major Population Groups

Tables 6 and 7 present coverage estimates as undercount rates for major population groups. The groups were selected for inclusion in this paper because they coincide with the variables used in modeling. Table 6 shows coverage estimates for major race and ethnic groups, owners and renters, and the nine age/sex categories used in modeling. These are the primary partitioning characteristics used in the post-stratification (although the modeling of race and ethnicity partitions uses seven mutually exclusive categories, while the table below includes persons with multiple race reports in more than one group, and includes all Hispanic persons in at least one race group).

Since each of the characteristics in Table 6 was used as a primary post-stratifier under the A.C.E. scheme, none of the differences between A.C.E. and CCM are significant. (They are not precisely equal, because the component rates are distributed across the entire census, whereas in Tables 9 and 10, they were applied only to the sample.) The characteristics that were either partitioned strictly (i.e. never combined in a post-stratum with any persons not sharing the characteristic) or given an enhanced importance under clustering show no significant differences with CCM. These are the race and Hispanic origin groups, tenure, and children's ages. The only significant differences are the Age/Sex groups aged 30-49 and 50+. Since the clustering allowed the ages to partition with each other, the 30-49 and 50+ group estimates compromised toward each other for both sexes, resulting in significant differences from CCM. Both age groups changed by about half a percent, in the direction of less undercount using clustering for ages 30-49 and the more undercount direction for 50+.

Table 6. Net Coverage for Primary Demographic Groups

		CCM		Post-stratification		Clustering	
	Census Count (×1000)	Percent Net Undercount (%)	SE (%)	Percent Net Undercount (%)	SE (%)	Percent Net Undercount (%)	SE (%)
National	300,703	-0.01	0.14	-0.14	0.13	-0.10	0.13
Race alone or in combination with one or more other races[1]							
White	225,547	-0.54	0.14	-0.64	0.13	-.0.62	0.13
Non-Hispanic White Alone	191,997	-0.83	0.15	-0.86	0.14	-0.84	0.15
Black	40.153	2.06	0.50	2.02	0.50	2.20	0.33
Asian	16,969	0.00	0.51	0.00	0.54	0.05	0.51
American Indian and Alaska Native	5,056	0.15	0.71	-0.31	0.72	-0.34	0.74
On Reservation	572	4.86	2.38	3.88	2.61	4.05	2.64
Am. Ind. Areas[2] Off Reservation	527	-3.86	2.98	-2.86	1.99	-3.26	2.06
Balance of the U.S.	3,959	-0.05	0.57	-0.61	0.56	-0.63	0.59
Native Hawaiian or Pacific Islander	1,189	1.02	2.07	0.39	2.05	-0.07	2.04
Some Other Race	21,448	1.63	0.32	1.18	0.33	1.17	0.29
Hispanic Origin	49,580	1.54	0.34	0.99	0.32	1.00	0.29
Tenure							
Owner	210,240	-0.57	0.12	-0.60	0.12	-0.57	0.13
Renter	99,463	1.09	0.29	0.78	0.31	0.85	0.30
Age/Sex							
0 to 4	20,158	0.72	0.40	0.39	0.39	0.50	0.38
5 to 9	20,315	-0.33	0.31	-0.46	0.31	-0.47	0.28
10 to 17	33,430	-0.97	0.29	-1.07	0.28	-0.95	0.28
18 to 29 Males	23,982	1.21	0.35	0.76	0.35	0.78	0.37
18 to 29 Females	23,912	-0.28	0.36	-0.49	0.35	-0.45	0.38
30 to 49 Males	40,256	3.57	0.20	3.47	0.20	2.98	0.19
30 to 49 Females	40,256	-0.42	0.21	-0.53	0.20	-1.02	0.19
50+ Males	41,815	-0.32	0.14	-0.32	0.14	0.17	0.14
50+ Females	51,950	-2.35	0.14	-2.35	0.14	-1.87	0.14

A positive estimate denotes a net undercount and a negative estimate denotes a net overcount.
Estimates are rounded for display.
The 2010 census population count excludes persons in group quarters and persons in Remote Alaska.
[1] A person can be included in more than one classification.
[2] American Indian Areas are lands considered (either wholly or partially) on an American Indian reservation/trust land, Oklahoma Tribal Statistical Area, Tribal Designated Statistical Area, or Alaska Native Village Statistical Area.
CCM results from Davis et. al. (2012).

4.3 Net Coverage for Other Groups

Table 7 presents estimates for additional population groups defined largely by CCM modeling characteristics. The first three, Region of the country, TEA group, and Census Participation Rate group, coincide with variables used as secondary stratifiers in the A.C.E. post-stratification (except that Update/Leave and Update/Enumerate are treated as one category in that scheme). The latter three, the Bilingual and Replacement Mailing area blocks, and household relationship types, are not used in the post-stratification.

Table 7. Net Coverage for Additional Groups

		CCM		Post-stratification		Clustering	
Net Coverage for other characteristics	Census Count (×1000)	Percent Net Undrcnt (%)	SE (%)	Percent Net Undrcnt (%)	SE (%)	Percent Net Undrcnt (%)	SE (%)
Census Region							
Northeast	53,618	-0.36	0.32	-0.06	0.22	-0.21	0.15
Midwest	65,156	-0.57	0.24	-0.18	0.18	-0.18	0.12
South	111,606	0.46	0.28	-0.11	0.20	-0.02	0.16
West	70,324	0.02	0.25	-0.19	0.19	-0.06	0.16
TEA Group							
Mailout/Mailback	278,553	0.02	0.14	-0.09	0.14	-0.02	0.14
Update/Leave	20,076	-1.37	0.66	-0.98	0.52	-1.70	0.40
Update/Enumerate	2,074	7.87	3.13	1.24	0.86	4.77	2.08
Census Participation Rate							
Low Rate	101,659	0.00	0.30	-0.08	0.31	-0.10	0.25
High Rate	199,044	-0.01	0.14	-0.26	0.14	-0.09	0.13
Bilingual Mailing Area							
Bilingual Mailing Areas	35,204	0.80	0.40	0.55	0.27	0.60	0.24
Balance of U.S.	265,499	0.12	0.15	-0.23	0.13	-0.19	0.13
Replacement Mailing Area							
Blanketed Mailing Areas	53,651	0.38	0.44	0.15	0.17	0.22	0.16
Targeted Mailing Areas	65,952	0.18	0.36	0.11	0.28	0.19	0.29
Balance of U.S.	181,100	-0.20	0.15	-0.32	0.12	-0.30	0.13
Nuclear Family Members	237,966	-0.32	0.14	-0.25	0.13	-0.38	0.14
Adult Children	24,036	-2.91	0.38	0.05	0.27	-2.92	0.36
Other Household Members	38,702	3.53	0.38	0.42	0.18	3.22	0.37

A positive estimate denotes a net undercount and a negative estimate denotes a net overcount.
Estimates are rounded for display.
The 2010 census population count excludes persons in group quarters and persons in Remote Alaska..

Of the characteristics used as A.C.E. stratifiers, the only ones that show a significant difference (under the A.C.E.-type post-stratification) compared with their CCM estimates are the Midwest and South regions. The Update/Enumerate TEA was combined with Update/Leave for A.C.E. post-stratification. Since the latter is much larger and shows an overcount for CCM, the Update/Enumerate was significantly underestimated under A.C.E. post-stratification. The A.C.E. post-stratification did not use the household relationship type at all, so the large CCM overcount for Adult Children and undercount for Other Household Members were both moved significantly in the direction of neutrality. The clustering partitioning differed significantly from CCM only for the South region. Overall, the clustering methodology captured important differences for characteristics that the A.C.E. scheme did not.

5. Conclusions

The twin objectives of coverage measurement, to estimate the coverage of demographic groups and to point out possible improvement for future censuses, presents a contradiction for the design of a post-stratification, because choices have to be made about which characteristics to prioritize and which to de-emphasize in the partitioning. A plan designed to capture the differential coverage of demographic groups will understate the difference among groups defined by operational characteristics, and vice versa. Logistic regression modeling avoids this compromise by assigning modeled values that capture the differential rates of all the characteristics used in the modeling. Its use should help the 2010 CCM program estimates to be useful for both purposes.

References

Bentley, M. (2008), “Specifications for Bilingual Form Distribution in the 2010 Census (Phase 1),” DSSD 2010 Decennial Census Memorandum Series #B-4.

Davis, P. and Mulligan, J. (2012), “2010 Census Coverage Measurement Estimation Report: Net Coverage for the Household Population in the United States,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-03.

Letourneau, E. (2010), “Specification to Identify Replacement Mailing Housing Units in the 2010 Census,” DSSD 2010 Decennial Census Memorandum Series #G-04-R1.

Mule, T. (2008), “2010 Census Coverage Measurement Estimation Methodology,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18.

Mulligan, J. (2012), “2010 Census Coverage Measurement: Description of Race/Hispanic Origin Domain,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-45.

Olson, D. (2012), “2010 Census Coverage Measurement Estimation Report: Aspects of Modeling,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-10.

Rothhaas, C., Bentley, M., Hill, J. M., and Lestina, F. (2011), “2010 Census: Bilingual Questionnaire Assessment Report,” DSSD 2010 CPEX Memorandum Series #C-01.

U.S. Census Bureau (2004), “Accuracy and Coverage Evaluation of Census 2000: Design and Methodology.”

Groupings under Post-stratifications

Table A-1: Grouping of Sample members into partitions with same characteristic

	E Sample			P Sample		
	Wgtd Pct	Post-strat	Clustered	Wgtd Pct	Post-strat	Clustered
AIAN on Reservation	0.19	100	100	0.19	100	100
AIAN off Reservation	0.61	100	100	0.69	100	100
Hispanic Origin	15.99	100	60.64	15.95	100	61.29
Black	12.02	100	47.71	11.62	100	46.96
Native Hawaiian or Pac. Is.	0.28	100	100	0.32	100	100
Asian	4.55	100	99.60	4.54	100	99.62
White	66.36	100	99.99	66.69	100	99.99
Owner	67.24	100	99.84	67.50	100	99.83
Renter	32.76	100	99.68	32.50	100	99.65
Child 0-4	6.59	99.87	99.75	6.77	99.89	99.73
5-9	6.71	99.87	97.01	6.90	99.89	97.04
10-17	11.10	100	98.15	11.25	100	98.14
Male 18-29	7.89	100	91.66	7.64	100	90.55
Female 18-29	7.76	100	96.57	7.80	100	96.51
Male 30-49	13.39	100	80.36	13.41	100	80.37
Female 30-49	13.89	100	83.86	14.07	100	84.19
Male 50+	15.20	100	85.09	14.93	100	85.20
Female 50+	17.47	100	87.77	17.22	100	87.67
Northeast	18.47	64.58	25.47	18.31	65.55	25.19
Midwest	21.03	67.95	33.30	21.16	68.57	33.36
South	37.57	70.78	49.57	37.28	70.75	49.32
West	22.93	58.07	28.34	23.26	58.69	28.76
Mailout/Mailback	92.15	98.84	95.78	92.17	98.49	95.67
Update/Leave	7.32	82.03	46.91	7.27	77.97	45.67
Update/Enumerate	0.53	27.53	84.78	0.56	26.79	79.72
Low Participation Rate	30.98	98.36	69.66	32.50	94.23	69.94
High Participation Rate	69.02	99.26	86.38	67.50	97.22	85.53
Bilingual Mailing Areas	10.62	34.35	25.48	89.50	92.29	91.27
Remaining Areas	89.38	92.20	91.10	10.50	34.21	25.57
Blanketed Mailing Areas	21.40	29.06	28.66	21.55	28.36	28.21
Targeted Mailing Areas	17.56	42.53	44.55	17.02	39.49	41.73
Remaining Areas	61.04	70.28	71.32	61.43	69.89	71.17
Nuclear Family Members	79.53	84.38	99.67	80.59	84.58	99.65
Adult Children	8.03	36.41	97.64	7.26	33.76	97.39
Other Household Members	12.45	17.60	97.91	12.15	16.58	97.80

Average Modeled Rates

Table B-1: Average Modeled CE Rates among E Sample persons

	Sample	Observed	Logistic	Post-stratified	Clustering
National	383,537	91.81	91.81	91.81	91.81
AIAN on Reservation	13,969	90.11	90.11	90.11	90.11
AIAN off Reservation	2,933	87.04	87.04	87.04	87.04
Hispanic Origin	64,301	90.58	90.58	90.58	90.16
Black	44,704	89.21	89.21	89.21	89.77
Native Hawaiian or Pac. Is.	4,047	87.03	87.03	87.03	87.03
Asian	20,823	91.73	91.73	91.73	91.71
White	232,760	92.66	92.66	92.66	92.66
Owner	241,557	93.69	93.69	93.69	93.69
Renter	141,980	87.95	87.95	87.95	87.96
Child 0-4	26,287	90.44	90.44	90.43	90.44
5-9	26,096	91.82	91.82	91.84	91.79
10-17	42,523	92.25	92.25	92.25	92.26
Male 18-29	31,815	87.08	87.08	87.08	87.18
Female 18-29	31,265	88.25	88.25	88.25	88.32
Male 30-49	51,041	92.08	92.08	92.08	91.90
Female 30-49	52,824	93.35	93.35	93.35	93.17
Male 50+	56,675	92.82	92.82	92.82	92.94
Female 50+	65,011	93.45	93.45	93.45	93.57
Northeast	65,673	91.94	91.94	92.12	91.82
Midwest	74,684	92.89	92.89	92.80	92.74
South	121,467	91.25	91.25	91.40	91.59
West	121,713	91.64	91.64	91.34	91.31
Mailout/Mailback	333,795	91.95	91.95	91.92	91.91
Update/Leave	32,085	90.05	90.05	90.51	90.58
Update/Enumerate	17,657	92.37	92.37	90.89	91.89
High Return Rate	131,311	89.26	89.10	89.32	89.46
Low Return Rate	252,226	92.96	93.03	92.93	92.87
Bilingual Mailing Areas	44,287	90.14	90.14	90.25	89.76
Remaining Areas	339,250	92.01	92.01	92.00	92.06
Blanketed Mailing Areas	82,093	90.95	90.95	91.39	91.36
Targeted Mailing Areas	71,105	88.70	88.70	89.74	89.43
Remaining Areas	230,339	93.01	93.01	92.56	92.66
Nuclear Family Members	299,051	93.01	93.01	92.27	93.00
Adult Children	31,300	87.77	87.77	89.63	87.74
Other Household Members	53,186	86.75	86.75	90.27	86.86

Table B-2: Average Modeled Match Rates among P Sample persons

	Sample	Observed	Logistic	Post-stratified	Clustering
National	355,812	91.07	91.07	91.07	91.07
AIAN on Reservation	12,761	83.55	83.55	83.55	83.55
AIAN off Reservation	3,224	87.80	87.80	87.80	87.80
Hispanic Origin	59,715	88.13	88.13	88.13	87.72
Black	39,677	87.01	87.01	87.01	87.57
Native Hawaiian or Pac. Is.	3,678	85.16	85.16	85.16	85.16
Asian	19,525	90.45	90.45	90.45	90.44
White	217,232	92.61	92.61	92.61	92.61
Owner	224,930	93.74	93.74	93.74	93.73
Renter	130,882	85.54	85.54	85.54	85.56
Child 0-4	25,372	88.30	88.30	88.30	88.31
5-9	25,073	90.45	90.45	90.45	90.50
10-17	40,144	91.49	91.49	91.49	91.45
Male 18-29	28,709	84.84	84.84	84.84	85.21
Female 18-29	29,340	86.41	86.41	86.41	86.58
Male 30-49	47,235	90.51	90.51	90.51	90.69
Female 30-49	49,393	92.03	92.03	92.03	92.41
Male 50+	51,239	93.73	93.73	93.73	93.38
Female 50+	59,307	94.36	94.36	94.36	93.98
Northeast	59,618	91.82	91.82	91.50	91.30
Midwest	70,604	93.06	93.06	92.46	92.30
South	111,607	89.95	89.95	90.59	90.79
West	113,983	90.48	90.48	90.24	90.23
Mailout/Mailback	309,691	91.20	91.20	91.17	91.14
Update/Leave	29,442	90.02	90.02	90.05	90.79
Update/Enumerate	16,679	84.02	84.02	87.54	84.03
High Return Rate	140,585	88.12	87.96	88.22	88.37
Low Return Rate	215,227	92.49	92.57	92.44	92.37
Bilingual Mailing Areas	40,352	88.48	88.48	88.32	87.91
Remaining Areas	315,460	91.38	91.38	91.40	91.44
Blanketed Mailing Areas	76,073	90.01	90.01	90.40	90.28
Targeted Mailing Areas	63,897	87.07	87.07	88.37	88.12
Remaining Areas	215,842	92.55	92.55	92.06	92.17
Nuclear Family Members	280,212	92.40	92.40	91.60	92.38
Adult Children	26,695	89.78	89.78	88.75	89.68
Other Household Members	48,905	83.07	83.07	88.93	83.25

Technical Notes on the Use of Clustering

Because the 2010 CCM estimation modeling uses some characteristics that are rare, but powerful when they occur, a method was needed to create post-strata using these rare events. The method chosen was to cluster them into groups of least 100 “nearest neighbors,” based on their Euclidean distance between the parameters of their modeling characteristics. For instance, two persons both living in Update/Enumerate operational areas, for that reason alone, were likely to be more similar to each other than to someone in a Mailout/Mailback neighborhood, even if that person shared traditional post-stratification characteristics like race and age.

The Euclidean distances used for the clustered partitioning were constructed from the parameters of a pair of logistic regression main effects models on the CCM E and P samples. Those parameters are listed in the table below. All characteristics except Participation Rate are categorical, and all categorical characteristics except Age are modeled by omitting one category from the parameterization to be used as a baseline. The parameter of the baseline category is implicitly estimated as the sum of the other parameters of the same category with the sign reversed. Age does not have a baseline because Children 0-17 were not classified by sex, so the Age parameters for that group implicitly absorb the parameter that contrasts sex against adults.

These parameters define 17 characteristics in both the E and P samples, including six for Age, totaling 34 levels. The individual parameters were assigned to the 34 variables. The Euclidean distance between any two persons is the sum of the squared difference between their parameters. For instance, two people with identical characteristics except that one lives in a Medium MSA and one in a Small MSA, would differ by the Euclidean distance between 0.0554 and 0.0404 (the Correct Enumeration parameters of those two MSA/TEA types), plus the distance from 0.2143 to 0.0597 (the Match rate parameters). This value is (0.0554 - 0.0404)^2 + (0.2143 - 0.0597)^2.

The assignment of clusters was performed by SAS Procedure FastClus, which partitions the input data set (in this case the E and P samples combined) into groups of nearest neighbors. Input options were set to permit the construction of not more than one thousand clusters, with a minimum cluster size of 200, in the expectation that no cluster would end up with fewer than 75 P-sample members, the minimum acceptable size of a post-stratum. FastClus created 382 clusters, none with fewer than 92 P-sample observations.

Due to expectations of the CCM environment, some interventions were deemed necessary or desirable in the cluster formation:

- AIAN on Reservations were clustered separately.
- Adult Males, Adult Females, and Children were clustered separately because the ratio of the DSE for the two adult sexes would become the basis for the estimation of correlation bias adjustment factors.
- The parameters associated with Race/Origin Domain and the Age categories of children were enhanced to improve their separation.

The two firm separations were achieved by setting the associated clustering parameters to arbitrarily high values of 10 or -10, which would prevent clustering with any other group. The enhanced separations were achieved by multiplying the associated parameters by five. This value was not arbitrarily. The FastClus software estimated the standard error among parameters for Tenure, Presence of Spouse, and Relationship to Householder in the range of about 5.0 to 7.5 and those groups were almost never clustered across. The standard error among the Race/Origin Domain and children's Age parameters was about one-fifth that much, and hence was multiplied by five in the hope of making group-crossing rare. This effort was largely successful, as the children's ages were clustered across less than three percent of the time; Race/Origin Domains were crossed less than one percent of the time, except for Hispanics and Blacks, whose parameters were very similar.

Appendix 1

Logistic Parameter Estimates used in Clustered Partitioning

Covariate	Correct Enum	Match
Intercept	0.4612	0.6082
Bilingual Mailing Area	-0.0259	-0.0352
Not a Bilingual Mailing Area		
Blanketed Mailing Areas	-0.0533	-0.0250
Targeted Mailing Areas	-0.0579	-0.0736
Not a Replacement Mailing Area		
AIAN on Reservation	0.0427	0.2005
AIAN off Reservation	-0.2100	-0.0027
Hispanic Origin	0.1516	0.0100
Black	0.1317	-0.0147
Native Hawaiian or Pac. Is.	-0.2019	-0.2717
Asian	0.0002	-0.0640
White		
Owner	0.2278	0.2790
Renter		
Male	-0.0691	-0.0831
Female		
Age 50+	0.0298	0.0488
Age 30-49	-0.0012	-0.0019
Age 18-29	-0.0520	-0.1147
Age 10-17	0.0349	0.0285
Age 05-09	-0.0007	-0.0002
Age 00-04 (no baseline)	-0.0056	0.0082
Spouse in Household	0.1636	0.1789
Other		
Northeast	0.0427	0.0633
Midwest	0.0121	0.1109
South	-0.0529	-0.1369
West		
Mailout, Medium MSA	0.0554	0.2143
Small MSA	0.0404	0.0597
non-MSA	0.0287	0.1199
Update/Leave, MSA	-0.2066	-0.0667
Update/Leave, non-MSA	-0.0561	0.0821
Update/Enumerate	0.0822	-0.6291
Mailout, Large MSA		
Participation Rate (continuous)	1.7634	1.2593
Adult Child	-0.1847	0.1498
Other Relation	-0.1219	-0.3445
Nuclear Family Member		

Census Operational Areas

Type of Enumeration Areas

The Type of Enumeration Area (TEA) is a classification of how the Census Bureau obtained addresses and conducted the census in an area. We provide estimates by combining six of the seven TEAs into three main categories. (The Remote Alaska TEA is out of scope.)

The first was "Mailout/Mailback," which included the Mailout/Mailback and the Military Mailout/Mailback TEAs. Questionnaires were delivered to housing units by mail and respondents were instructed to return the form by mail.

The second category was the "Update/Leave," which included the Update/Leave and the Urban Update/Leave TEAs. A census worker updated the address list and delivered questionnaires to each address that was on the updated address list. Respondents were instructed to return the form by mail.

The third was the "Update/Enumerate," which included the Remote Update/Enumerate and the Update/Enumerate TEAs. A census enumerator updated the address list and conducted the enumeration at each housing unit on the updated address list.

Bilingual Mailing Areas

For the 2010 Census, the Census Bureau mailed a bilingual (English and Spanish) census questionnaire to housing units in select areas that could require Spanish language assistance to complete their census form. For more information on bilingual mailing, see Bentley (2008) or Rothhaas et al. (2011). We estimated coverage for the areas that received the bilingual questionnaire versus the remainder of the country.

Replacement Mailing Areas

For the 2010 Census, the Census Bureau mailed a replacement mailing package to some housing units in Mailout/Mailback areas of the country that had low mail response in Census 2000. Areas with low response in Census 2000 had a blanketed distribution where all housing units received a replacement mailing. For areas with mid-range response in 2000, only nonresponding housing units received a replacement mailing; this is referred to as targeted distribution. The balance of the United States did not receive a replacement questionnaire in the mail. We provided separate estimates for the two types of replacement mailing areas (blanketed and targeted) and the balance of the United States. For more information on the replacement mailing areas and the official counts, see Letourneau (2010).